A New View of Parameter Search

Richard M Young

Psychology Department
University of Hertfordshire

Talk presented at Act-R Workshop
Carnegie Mellon University
25-27 July 2003

Outline of Talk

OUTLINE

- A parameter space landscape
- New view of parameter search
- Prediction patterns and structural constraints
- Implications for model fitting and parameter estimation

AIMS

- Not tell you anything you don't already know
  — at least, implicitly
- But hope to change forever the way you think about it!

A Parameter Space Landscape

- Based on an email sent to Act-R mailing list, 17 Jan 2001
- The "slow Kendler" model
  — Niels Taatgen's model of Kendler & Kendler (1959) data, as described in 1998 Act-R book
- The data to be fit

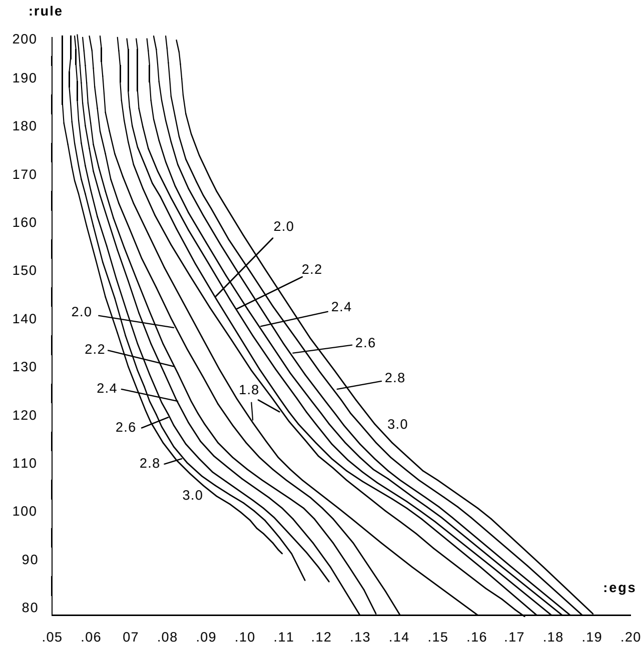
	Initial	Trial
Reversal	7.3	24.4
Extra-dimensional	7.3	9.0

- The model parameters to be adjusted for best fit
  a) :egs — expected gain noise
  b) :eventual-successes on newly learned rules (':rule')
- In the deviation space defined by the two parameters, informal search reveals a more-or-less well behaved bowl-shaped region, with a more-or-less well defined minimum.
  ("deviation space" — *i.e. a map of the deviation between model predictions and data*)

The Valley (or 'Gorge')

- More careful exploration reveals that a valley leaves the bowl
  — heads NW
  — then curves round to head almost due North
- The valley
  — becomes very straight
  — almost parallel to the :rule axis
  — very steep-sided
  — very narrow, e.g. the dev=1.8 contour only 0.005 of an :egs unit wide (at :rule = 200)
  — very level, e.g. the dev=1.8 contour on its floor along its whole length.
- Bad news for parameter estimation
  — no single point where parameter values give best fit
  — instead there's a whole *contour* (along the floor of the valley) of "best" values.
- This kind of feature is part of the (connectionist) lore, but surprising in this model: Why does it arise?
- Can't show picture of whole space
  — need different scales in different regions of the space
  — working on it!

Part of the "Slow Kendler" Parameter Space



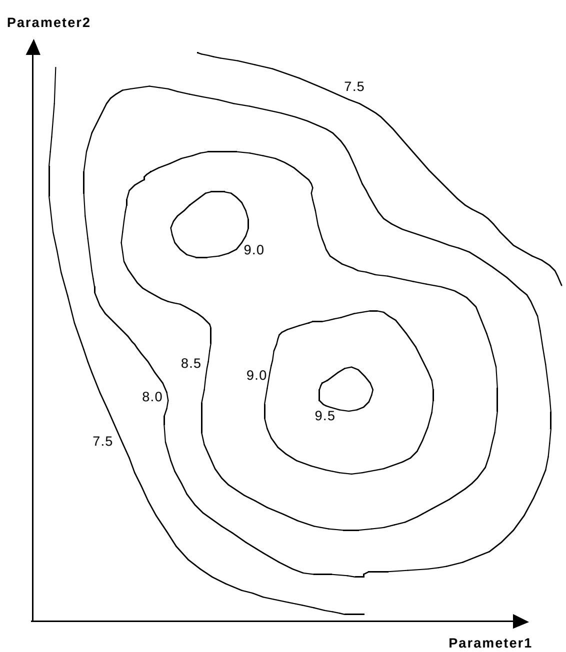
A New View: Prediction Landscape

- Stand back, and take a fresh look at what we're doing when we search a 2D (multi-D) parameter space to fit model predictions to empirical data.
- Imagine a 2-parameter space.
- Suppose the model predicts a single measurement, which can be compared with (corresponding measurement) in the data.
  — Obviously silly to expect to be able to determine 2 parameter values against a single point prediction … but bear with me!

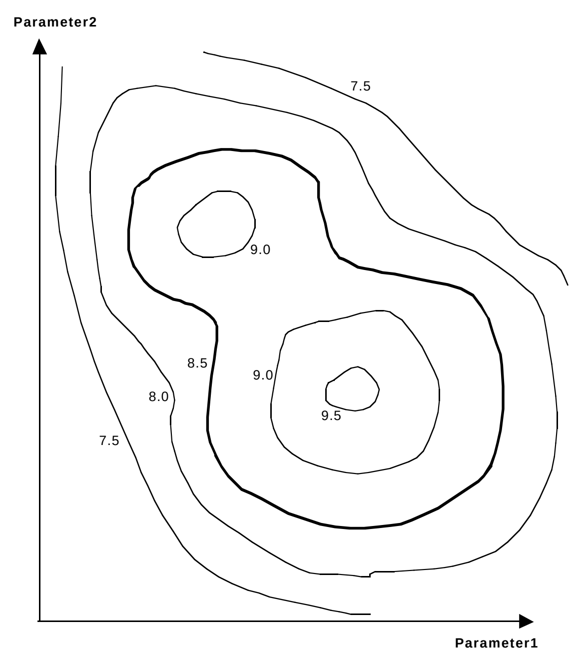
- Note: This is a predicted-**value** landscape, not a **deviation** landscape.

- Some particular level in the landscape corresponds to the actual value of the measurement in the data. Say, data has value 8.5. The can identify the corresponding contour.
- So the model 'fits' the data anywhere along the contour
  — i.e. for any of the pairs of parameter values on the contour.

Hypothetical "Value" Landscape



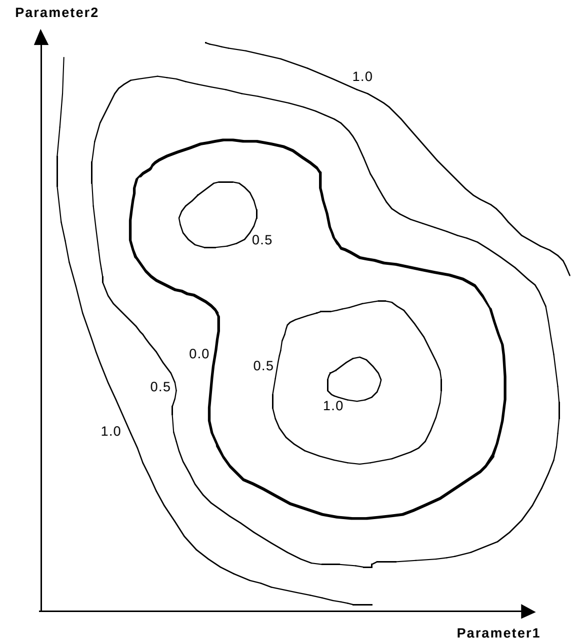
Hypothetical "Value" Landscape with Data Contour Marked



Fold to Get a Deviation Landscape

- How do we transform this *value* landscape into a *deviation* landscape?
- Whereas the value landscape can be above or below the best-fit contour, in the deviation landscape one only be *above* it.
- If think of the value landscape as being shaped from paper or cloth, then trick is to *fold* it along the best-fit contour
  — so that parts that were above stay above, and parts that were below are now also above.
- Reflect the parts below in a horizontal plane at the level of the best-fit contour
  — get a crease along the best-fit contour
- Analogy with clear but reflecting lake.
- The contour lines stay the same shape, but their numbers change.
- **Note** that the deviation landscape is "ambiguous"
  — loses information relative to value landscape

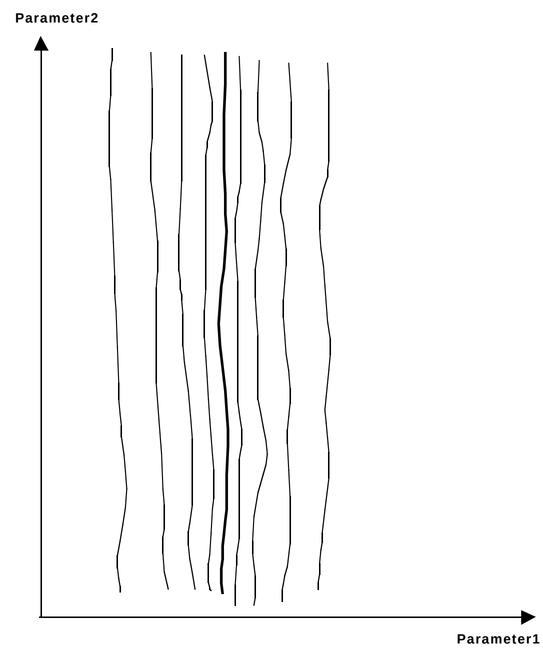
"Deviation" Landscape



Origin of Long Narrow Valleys

- Suppose that in some region of the parameter space, the value predicted by the model is very sensitive to one of the parameters (say, parameter1).
- Picture can be read as either value or deviation landscape.
- In the value landscape, the land locally is something like a plane surface, tilted steeply parallel to the parameter2 axis.
- In the deviation landscape, we get a narrow, steep-sided V-shaped valley parallel to the parameter2 axis.
- Note how the properties of the valley follow directly from the topography:
  — level-bottomed, because lies along the best-fit contour
  — long, because along length of best-fit contour
  — narrow & steep-sided, because sensitive to parameter1
  — straight & nearly parallel to parameter2 axis, because changes due to parameter1 dominate those of parameter2

Value/Deviation Map of Long Narrow Valley



Multiple Values

- To make realistic, now need to extend the analysis to case where model predicts multiple values. Suppose the model predicts N values corresponding to measurements in the data.
- In general (in the mathematical sense), if the N values are independent functions of the parameters, the story falls apart. But this is just about never the case.
- Instead, the model predictions form a definite *pattern*
  — usually due to the structure of the model
  — holds approximately constant, or changes only slowly with changes in parameters
- The slow Kendler model predicts essentially three values:
  — # trials to initial learning (~ 7)
  — # trials to re-learning after reversal shift (~ 24)
  — # trails to re-learning after extra-dimensional shift (~ 9)
- Although the actual numbers depend on the parameter settings, their relationships stay fairly constant. Consequences of the structure of the model are that
  — reversal learning takes around 2.5-3 times initial learning
  — never have extra-dimensional take longer than reversal

Patterns and Levels

- So, effect of changing parameter settings is to change the level of predictions as a whole, but to leave their pattern relatively unchanged.
- For example, for Slow Kendler, define
  — the *level* as mean of the three quantities (~ 13)
  — the *pattern* as differences from the mean {–6, –4, +10}
- In other cases, may need to be more ingenious in defining an appropriate level and approximately invariant pattern.
- The analysis now holds for real models, which predict multiple data points, with these small adaptations:
  — the story applies to the *level* of the model predictions, rather than to just a single model prediction
  — the minimum deviation is no longer zero, because the *pattern* of the model predictions won't coincide exactly with the *pattern* of the data.
- **Note** that even if the above assumptions (about constancy of pattern, etc.) aren't met across the whole parameter space, will tend to hold near the best-fit values
  — which is all that matters

This Analysis Undermines the Basis for Parameter-Fitting

- Should not underestimate how radical this story is.
- We've cut the ground away from the assumption normally made, that *one can pin down parameter settings by fitting the model to data.*
- Usually there will not be a best-fitting point in the parameter space. Usually there will be a whole contour (or in higher Ds, a subspace or manifold or whatever) with a more-or-less equally good fit to the data.
- (Deal with response of fitting multiple data sets, e.g. latency and errors, simultaneously.)
- (The fixed structure of the model, i.e. the contents of the productions, pins down many degrees of freedom.)

Implications for Model Fitting — 1

- Before accepting the results of optimised parameter settings, always explore, map, and understand the topography of the parameter space.
  — don't trust optimum fit if only slightly better than others
  — if there's a best-fitting contour rather than a single point, don't commit to a single set of values.
- Better to try standardised parameter settings and check that get decent fit, rather than choosing settings for a data set by optimising fit.
  — reinforces this idea, which is around anyway;
  — optimising the fit actually works *against* getting parameter setting stable over a range of experiments.
- To understand your model and its predictions, better to plot and study the ***value*** landscape rather than the ***deviation*** landscape.
  — the deviation landscape is "ambiguous", and can make the picture harder to interpret;
  — to do this, need to define an appropriate measure of "level" of the model predictions.

Implications for Model Fitting — 2

- We can't remove the arbitrariness of picking one point on a contour by using a second dataset and taking the intersection of the contours.
  — picking a common point just means that the arbitrariness is shared.

- The commonly held belief that optimised settings are safe if # data points >> # parameters, is *wrong*.
  — Roberts & Pashler (2000) are right about this;
  — to the extent it is true, apply with extreme caution;
  — whole issue needs to be re-visited, and understood better.